DESCRIPTION

DJ
#3 DE920000[illegible]-02
Records Paper / Microfiche Paper

METHOD AND APPARATUS FOR THE AUTOMATIC SEPARATING AND INDEXING OF MULTI-SPEAKER CONVERSATIONS

jc872 U.S. PTO
09/997957
11/30/01

Background of the Invention

The invention generally relates to the field of digital audio processing and more specifically to a method and apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction. The invention further relates to a multi-user speech recognition or voice control system.

Business transactions are increasingly conducted by way of telephone conversation. Exemplarily it is referred to audio logs of call center dialogues which have to be accessed in order to locate specific transactions. Another example are logs which are stored on audio tapes and can be accessed by scanning corresponding tape archives.

Beyond that it is to be expected that in the future many transactions like teleshopping or telebanking will be handled by automatic transaction systems using text to speech synthesis to communicate with a customer. Another substantial and still growing amount of transactions is the field of telephone conversation which takes place between two human individuals, in particular two individuals speaking different languages.

A particular field of transactions is transactions that are legally binding. It is current practice to record the underlying interactions on audio tapes to have a log of each interaction. For legal reasons, in cases where both parties disagree about an intended transaction at a later date, these logs can be used as a proof instrument. Nowadays such tapes are labeled with a date information and a customer or employee identifier. This makes the task of locating and indexing an audio log of a specific transaction an extraordinary effort.

Prior efforts to automize the indexing of such audio material, e.g. using prior art speech recognition technology, failed due to the large variability of speech styles and dialects of the human individuals engaged in those interactions.

Another application field is multi-user speech recognition systems (SRSs) where two or more speakers are located in the same room, e.g. a typical mixed conversations during personal meetings or the like which shall be protocolled using SRS technology. Another similar situation is command language used in an aircraft cockpit where the pilot and the co-pilot operate the aircraft via voice control. As modern SRSs have to be trained for different users, these systems so far are not able to automatically switch between the different speakers.

Summary of the Invention

It is therefore an object of the present invention to provide a method and apparatus which allow to simplify the aforementioned processing of a continuous audio stream containing human speech.

It is another object to provide such a method and apparatus which allow for an automized processing of an audio stream incoming in real-time or being stored on a storage media.

It is yet another object to provide such a method and apparatus which reduce the costs and time efforts for locating specific transactions or speaker-related audio segments in a continuous audio stream.

The above objects are solved by the features of the independent claims. Advantageous embodiments are subject matter of the subclaims.

The idea underlying the invention is to locate segments in a continuous audio stream where a change-over to at least one predefined speaker occurs and to apply speech recognition or voice control techniques only to those audio segments belonging to the predefined speakers.

In view of the common practice in commercial or business transaction-related conversations or dialogues, in order to avoid miscommunication, the essential information to identify a customer, employee or the like is obtained as customer name or account number uttered and repeated at the beginning of a dialogue and thus the proposed mechanism is able to capture all the essential information necessary to identify and transcribe the audio information related to the particular underlying transaction.

More particularly, the invention proposes to apply known speaker recognition techniques to conversations between a well-known speaker and a multitude of unknown speakers and thereby allows to transcribe only the utterances of the well-known speaker as index and summary of the dialogues.

It is noteworthy that the two steps of detecting at least one speaker change in the continuous audio stream and of performing a speaker recognition for the audio stream at least after an allegedly detected speaker change can be performed in an arbitrary order. Performing a speaker change detection prior to performing a speaker recognition has the advantage that the processing resources and time consuming mechanism of speaker recognition must only be executed if a speaker change is detected wherein the speaker change detection process is much less consuming resources than the speaker recognition. On the other hand, executing both steps in the reverse order has the advantage that the speaker change can be detected using the results of the speaker recognition and must not be implemented as an independent step thus simplifying the entire

According to a first alternative of the invention, a real-time incoming continuous audio stream, e.g. speech that is going to be transcribed by a speech recognizer or an incoming telephone call, is scanned in order to detect a speaker change. Further it is analyzed if the detected audio segment(s) is belonging to a predetermined or preselected speaker wherein only those audio segments belonging to the predetermined speaker(s) are transcribed e.g. into plain text by way of speech recognition.

As a second alternative, a continuous audio stream, e.g. a telephone call or the like, is first recorded on a record media like a magnetic tape, CD-ROM or a computer hard disk drive (HDD) and the recorded audio stream is scanned in order to detect audio segments belonging to a predefined speaker. These audio segments are then indexed and only the indexed segments are transcribed into spoken or written language later on. Thus a particular human speech-based transaction can be found in a large, unstructured storage media like a magnetic tape.

In a third alternative, the invention is used to enable speaker-triggered speech or voice recognition in a multi-user speech recognition or voice control environment providing, for each user, a different speaker model and optionally a different dictionary or vocabulary of words already known or trained by the speech or voice recognition system. In such an environment it is necessary to switch between different dictionaries when a first user has stopped utterance and a second user is going to start his utterance. Hereby a real-time continuous audio stream has to be processed in order to distinguish between utterances of the different users.

It should be noted that use of the invention is by no means limited to the above mentioned application fields and thus can be used or implemented, for instance, in a voice activation control system of an automobile or aircraft or the like. It can also be used to separate background speech signals in order to filter those signals from a currently interesting speech signal or utterance, e.g. in a scenario where two or more people are staying in the neighborhood or at least within an audible distance, each of them using a speech recognition or dictating system or a voice control system.

Brief Description of the Drawings

In the following the invention will be described in more detail referring to the accompanying drawings from which further features and advantages will become evident. In the drawings

Fig. 1a is a flow diagram which illustrates the basic features and steps of the method according to the invention;

Fig. 1b is another flow diagram which illustrates a more detailed embodiment of the invention;

Fig. 2 is a block diagram depicting the basic components of a first embodiment of the apparatus according to the invention;

Fig. 3 is another block diagram depicting a second embodiment of the apparatus according to the invention; and

Fig. 4 shows an example of a log file encoded using XML markup language in accordance with the invention.

Detailed Description of the Drawings

Fig. 1a shows the basic steps of a routine processing a continuous audio stream in accordance with the invention. After the routine is started 10 and the audio stream is digitized (not shown here) the digitized audio stream is analyzed in order to locate speaker changes 20. A lot of speaker change and speaker detection algorithms are known in the literature. For a comparison of techniques see for example F. Bimbot et al., Second-Order Statistical Measures for Text-Independent Speaker Identification, Speech Communication, Vol. 17, p.177-192, 1995. Hereby, for instance, the audio stream can be analyzed at frequency bands in order to derive characteristic features for different speakers. For a speaker change detection such feature vectors may be subjected to classical change detection techniques as described in the textbook by M. Basseville and Igor V. Nikiforov, Detection of Abrupt Changes: Theory and Applications, Prentice Hall Englewood Cliffs 1993, whereas for speaker identification the features are matched against a database of known speakers (S. Furui, An Overview of Speaker Recognition Technology, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, p.1-9, Martigny 1994).

If a speaker change is detected, for at least part of the following audio stream a speaker recognition is performed 30. Otherwise the speaker change detection is repeated until a speaker change is detected at all. After the speaker recognition 30 is finished it is checked 40 whether the recognized speaker is equal to a predetermined or preselected speaker or alternatively whether the speaker is recognized as a known speaker at all. If so at least the above mentioned part of the audio stream is transcribed, e.g. into plain text by means of a known speech recognition technique.

Now referring to the flow diagram depicted in Fig. 1b, a continuous audio signal 100, either recorded by means of an analog storage media or provided real-time, is first digitized 105. The digitized audio data are then scanned 110 whereby it is checked during loop 115 whether a speaker change occurs 120 and whether the detected new speaker is identical with a predefined or known speaker. The latter step is performed by means of speaker recognition 130 using

It is emphasized that the steps of detecting a speaker change 120 and performing a speaker recognition 130 can be alternatively performed in the reverse directon wherein the results of the speaker recognition 130 themselves can be used in order to detect speaker changes 120 thus simplifying the above described approach.

If the speaker change detection 120 reveals that a speaker change has occurred, the current time is taken 125 and protocolled e.g. in a log file. Having performed the speaker recognition 130 it is checked 135 whether the recognized speaker is identical with a predefined speaker. If true, the audio segments starting with the detected speaker changes are indexed 140 by using the protocolled time 125.

The scanning of the audio stream is continued 150 until the entire audio stream is scanned through and analyzed in the above described manner. Having finished the scan, only for the segments corresponding to selected speakers a speech recognition procedure, as known in the prior art, is performed 160. Hereby, in a preceding step 155, a speaker-related voice tract model and/or dictionary for the recognized speaker (step 130) is selected wherein the speech recognition 160 is performed based on that dictionary.

It is further noted that the steps 110 and 150 are optional and related to a scenario where an audio stream stored on a data carrier is scanned offline in order to perform the method according to the invention. Without these steps the mechanism can be performed for a real-time audio stream like a speech signal incoming in a speech or voice recognition system.

Using a time base generator (step 102) as an external time reference for enabling writing of the time tags (step 140) is also optional and needed only in cases where the original audio signal does not comprise timing information.

The described method advantageously enables to perform speech recognition only for audio segments in a continuous audio stream which have been uttered by a given speaker.

Fig. 2 depicts a first embodiment of an apparatus according to the invention. In this embodiment, the continuous audio stream is recorded on a tape storage 200. First the audio stream is digitized by means of a prior art digitizer 210 particularly revealing digitized timer information 220 for the audio stream. In addition, the digitized audio stream is searched for speech/non-speech boundaries by means of an appropriate detector 230 also well-known in the prior art. The non-speech detector 230 delivers first candidates of speaker-change boundaries in form of first audio segments.

For these audio segments an utterance analysis is performed by means of an utterance analyzer & change detector 240. The audio

stream is analyzed by an utterance analyzer which scans through the audio stream in order to gather speaker-specific audio features. For instance, the utterance analyzer can be implemented as a spectrum analyzer which takes information in the neighborhood of frequency bands which are characteristic for different speakers. The analyzed utterance signal is forwarded to an utterance changed detector which detects speaker changes. If an utterance or speaker change is detected by detector 240, the time of the speaker change is taken from the timer information provided by the digitizer 210, or an external timer, and written to a log file 255 stored in a database 260 by means of an indexer 250. It should be noted, that in many SRS systems, the utterance analyzer is already an integrated part of the SRS (e.g. P.S. Gopalakrishnan et al, Acoustic models used in the IBM System for the ARPA HUB4 task, Proc. of the Speech Recognition Workshop, ARPA, 1996).

For most applications, the time information alternatively can be taken from the clock of a computer system or a dedicated hardware that is used to perform the speaker recognition. In cases where a higher precision is needed for the timing information, e.g. in an automatic logging or indexing of air control dialogues, the time can be taken from an external time reference that is merged with the audio stream during the digitization step.

Taking the logged index information together with the digitized audio stream provided by the digitizer 210, a speech recognition system (SRS) 270 as known in the prior art can perform a speech recognition procedure on the audio stream, but solely for the indexed audio segments.

It should be noted that the system described herein before processes audio data digitized by prior art technology. In e.g. a call center environment, such data are usually collected from the telephone set or the head set of an operator. For logging and archiving the digitized data stream is stored in a file, either on a call-by-call or shift-by-shift basis. The same digitized audio stream is then passed through the described speaker recognition system that computes features allowing the identification of individual speakers.

Now referring to Fig. 3, a second embodiment of the apparatus according to the invention is described. A real-time audio stream is input to a microphone 300 and digitized by means of a digitizer 310. The digitized audio stream is input to an utterance analyzer & change detector 320 in order to detect speaker changes as described above. A speech recognition system (SRS) 330 has implemented a speaker model and/or dictionary change utility 340 which has access to different speaker-trained data 360, 370 stored in a database 350. Dependent on allegedly detected speaker changes, the dictionary change utility 340 can interchange between the different models 350, 370 thus providing an optimized multi-user SRS.

An example of a log file encoded using XML markup language in accordance with the invention is depicted in Fig. 4. The shown call

center scenario starts with an incoming customer call 400 and a welcome text 410 spoken by an operator of the call center. The operator is assumed to be a preselected speaker with a known speaker ID which is ~s0127" in the present example. Thus the start time and the end time of the welcome text 410 are marked with corresponding tags 420, respectively. The customer not being a preselected speaker with an ID is responding to the welcome text 410 and the audio signal tagged with the corresponding start time and end time accordingly but storing that the speaker ID is ~unknown~. Next the operator asks the customer for the customer number 440 wherein the audio signal is tagged again 450 with the known speaker ID. These steps are continued accordingly until the end of the call wherein in step 460 the operator repeats the customer number named by the customer in the preceding step and confirms the correct database entry of the customer ~and the address is Helga Mustermann the ...~. At the end of the call the audio signal is tagged with the endcall time 470.

It should further be noted that the above described method and apparatus can either be implemented in hardware, software or a combination thereof.

C L A I M S

1. A method of processing a continuous audio stream containing human speech related to at least one particular transaction, comprising the steps of:

digitizing the continuous audio stream;

detecting a speaker change in the digitized audio stream;

performing a speaker recognition if a speaker change is detected;

transcribing at least part of the continuous audio stream if a predetermined speaker is recognized.

2. A method of processing a continuous audio stream containing human speech related to at least one particular transaction, comprising the steps of:

digitizing the continuous audio stream;

detecting a speaker change in the digitized audio stream;

performing a speaker recognition if a speaker change is detected;

indexing the audio stream with respect to the detected speaker change if a predetermined speaker is recognized.

3. Method according to claim 1 or 2, comprising the further step of protocolling time information for detected speaker changes.

4. Method according to any of the preceding claims, wherein the step of detecting a speaker change and/or the step of performing a speaker recognition is/are preceded by the further step of detecting non-speech boundaries between continuous speech segments.

5. Method according to any of the preceding claims, wherein the step of detecting a speaker change is accomplished by use of at least one characteristic audio feature, in particular features derived from the spectrum of the audio signal.

6. Method according to claim 1 or 2, wherein the step of performing a speaker recognition involves the particular steps of calculating a speaker signature from the audio stream and comparing the calculated speaker signature with at least one known speaker signature.

7. Method according to any of the preceding claims for use in a speech recognition or voice control system comprising at least two speaker-specific speaker models and/or dictionaries, wherein interchanging between the at least two speaker-specific dictionaries dependent on the detected speaker change and the corresponding recognized speaker.

8. Apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction, comprising:

means for predetermining at least one speaker;

means for detecting speaker changes in the audio stream;

means for recognizing the predetermined speaker in the audio stream;

means for initiating transcription of at least part of the audio stream in case of a detected speaker change and a recognized predetermined speaker.

9. Apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction, comprising:

means for predetermining at least one speaker;

means for detecting speaker changes in the audio stream;

means for recognizing the predetermined speaker in the audio stream;

means for indexing the audio stream dependent on a detected speaker change and a recognized predetermined speaker.

10. Apparatus according to claim 8 or 9, further comprising

means for detecting non-speech boundaries between continuous speech segments.

11. Apparatus according to any of claims 8 to 10, further comprising

means for automatically scanning a continuous audio record, in particular a continuous audio stream recorded on a data or a signal carrier, and for detecting speaker changes in the continuous audio record.

12. Apparatus according to any of claims 8 to 11, further comprising

means for continuously monitoring a real-time continuous audio stream and performing the steps of claim 1 or 2.

13. Apparatus according to any of claims 8 to 12, further comprising log means for protocolling time information for the at least one detected speaker change.

14. Apparatus according to any of claims 8 to 13, comprising means for marking at least the beginning of a detected speech segment related to a predetermined speaker.

15. Apparatus according to any of claims 8 to 14, comprising data base means for storing speech signatures for at least two speakers.

16. Speech recognition or voice control system processing an incoming audio stream and having at least two speaker models and/or speaker-specific dictionaries, comprising

means for detecting a speaker change in the incoming audio stream;

means for gathering speaker-specific information and for comparing the gathered speaker-specific information with corresponding speaker-specific information of at least one predetermined speaker thus recognizing the at least one predetermined speaker;

means for interchanging between the at least two speaker-specific dictionaries dependent on the detected speaker change and the corresponding recognized speaker.

17. A data processing program for execution in a data processing system comprising software code portions for performing a method according to any of claims 1 to 7 when said program is run on said computer.

18. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform a method according to any claims 1 to 7 when said program is run on said computer.

A B S T R A C T

Disclosed are a method and apparatus for processing a continuous audio stream containing human speech in order to locate a particular speech-based transaction in the audio stream, applying both known speaker recognition and speech recognition techniques. Hereby it is enabled that only the utterances of a particular predetermined speaker are transcribed thus providing an index and a summary of the underlying dialogue(s).

In a first scenario, an incoming audio stream, e.g. a speech call from outside, is scanned in order to detect audio segments of the predetermined speaker. These audio segments are then indexed and only the indexed segments are transcribed into spoken or written language. Thus an already occurred specific transaction can be found on an endless storage media like a magnetic tape. The proposed mechanism thus makes the task of locating an audio log of a specific transaction a much more less effort.

In a second scenario, two or more speakers located in one room are using a multi-user speech recognition system (SRS). For each user there exists, a different speaker model and optionally a different dictionary or vocabulary of words already known or trained by the speech or voice recognition system. In such an environment, the invention allows to switch between different dictionaries when a first user has stopped utterance and a second user is going to start his utterance.
(Fig. 1B)

Europäisches Patentamt
European Patent Office
Office européen des brevets

Jc872 U.S. PTO
09/997957
11/30/01

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

00127335.8

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE, 14/05/01
LA HAYE, LE

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 00127335.8
Demande n°:

Anmeldetag:
Date of filing: 30/11/00
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
International Business Machines Corporation
Armonk, N.Y. 10504
UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
Method and apparatus for the automatic separating and indexing of multi-speaker conversations

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE/TR
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

30.11.2000 EP00127335.8 DESC DE920000055

EPO - Munich 3 30. Nov. 2000

D E S C R I P T I O N

METHOD AND APPARATUS FOR THE AUTOMATIC SEPARATING AND INDEXING OF MULTI-SPEAKER CONVERSATIONS

Background of the Invention

The invention generally relates to the field of digital audio processing and more specifically to a method and apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction. The invention further relates to a multi-user speech recognition or voice control system.

Business transactions are increasingly conducted by way of telephone conversation. Exemplarily it is referred to audio logs of call center dialogues which have to be accessed in order to locate specific transactions. Another example are logs which are stored on audio tapes and can be accessed by scanning corresponding tape archives.

Beyond that it is to be expected that in the future many transactions like teleshopping or telebanking will be handled by automatic transaction systems using text to speech synthesis to communicate with a customer. Another substantial and still growing amount of transactions is the field of telephone conversation which takes place between two human individuals, in particular two individuals speaking different languages.

A particular field of transactions is transactions that are legally binding. It is current practice to record the underlying interactions on audio tapes to have a log of each interaction. For legal reasons, in cases where both parties disagree about an intended transaction at a later date, these logs can be used as a proof instrument. Nowadays such tapes are labeled with a date

information and a customer or employee identifier. This makes the task of locating and indexing an audio log of a specific transaction an extraordinary effort.

Prior efforts to automize the indexing of such audio material, e.g. using prior art speech recognition technology, failed due to the large variability of speech styles and dialects of the human individuals engaged in those interactions.

Another application field is multi-user speech recognition systems (SRSs) where two or more speakers are located in the same room, e.g. a typical mixed conversations during personal meetings or the like which shall be protocolled using SRS technology. Another similar situation is command language used in an aircraft cockpit where the pilot and the co-pilot operate the aircraft via voice control. As modern SRSs have to be trained for different users, these systems so far are not able to automatically switch between the different speakers.

Summary of the Invention

It is therefore an object of the present invention to provide a method and apparatus which allow to simplify the aforementioned processing of a continuous audio stream containing human speech.

It is another object to provide such a method and apparatus which allow for an automized processing of an audio stream incoming in real-time or being stored on a storage media.

It is yet another object to provide such a method and apparatus which reduce the costs and time efforts for locating specific transactions or speaker-related audio segments in a continuous audio stream.

The above objects are solved by the features of the independent

claims. Advantageous embodiments are subject matter of the subclaims.

The idea underlying the invention is to locate segments in a continuous audio stream where a change-over to at least one predefined speaker occurs and to apply speech recognition or voice control techniques only to those audio segments belonging to the predefined speakers.

In view of the common practice in commercial or business transaction-related conversations or dialogues, in order to avoid miscommunication, the essential information to identify a customer, employee or the like is obtained as customer name or account number uttered and repeated at the beginning of a dialogue and thus the proposed mechanism is able to capture all the essential information necessary to identify and transcribe the audio information related to the particular underlying transaction.

More particularly, the invention proposes to apply known speaker recognition techniques to conversations between a well-known speaker and a multitude of unknown speakers and thereby allows to transcribe only the utterances of the well-known speaker as index and summary of the dialogues.

It is noteworthy that the two steps of detecting at least one speaker change in the continuous audio stream and of performing a speaker recognition for the audio stream at least after an allegedly detected speaker change can be performed in an arbitrary order. Performing a speaker change detection prior to performing a speaker recognition has the advantage that the processing resources and time consuming mechanism of speaker recognition must only be executed if a speaker change is detected wherein the speaker change detection process is much less consuming resources than the speaker recognition. On the other hand, executing both steps in the reverse order has the

advantage that the speaker change can be detected using the results of the speaker recognition and must not be implemented as an independent step thus simplifying the entire mechanism.

According to a first alternative of the invention, a real-time incoming continuous audio stream, e.g. speech that is going to be transcribed by a speech recognizer or an incoming telephone call, is scanned in order to detect a speaker change. Further it is analyzed if the detected audio segment(s) is belonging to a predetermined or preselected speaker wherein only those audio segments belonging to the predetermined speaker(s) are transcribed e.g. into plain text by way of speech recognition.

As a second alternative, a continuous audio stream, e.g. a telephone call or the like, is first recorded on a record media like a magnetic tape, CD-ROM or a computer hard disk drive (HDD) and the recorded audio stream is scanned in order to detect audio segments belonging to a predefined speaker. These audio segments are then indexed and only the indexed segments are transcribed into spoken or written language later on. Thus a particular human speech-based transaction can be found in a large, unstructured storage media like a magnetic tape.

In a third alternative, the invention is used to enable speaker-triggered speech or voice recognition in a multi-user speech recognition or voice control environment providing, for each user, a different speaker model and optionally a different dictionary or vocabulary of words already known or trained by the speech or voice recognition system. In such an environment it is necessary to switch between different dictionaries when a first user has stopped utterance and a second user is going to start his utterance. Hereby a real-time continuous audio stream has to be processed in order to distinguish between utterances of the different users.

It should be noted that use of the invention is by no means

limited to the above mentioned application fields and thus can be used or implemented, for instance, in a voice activation control system of an automobile or aircraft or the like. It can also be used to separate background speech signals in order to filter those signals from a currently interesting speech signal or utterance, e.g. in a scenario where two or more people are staying in the neighborhood or at least within an audible distance, each of them using a speech recognition or dictating system or a voice control system.

Brief Description of the Drawings

In the following the invention will be described in more detail referring to the accompanying drawings from which further features and advantages will become evident. In the drawings

Fig. 1a is a flow diagram which illustrates the basic features and steps of the method according to the invention;

Fig. 1b is another flow diagram which illustrates a more detailed embodiment of the invention;

Fig. 2 is a block diagram depicting the basic components of a first embodiment of the apparatus according to the invention;

Fig. 3 is another block diagram depicting a second embodiment of the apparatus according to the invention; and

Fig. 4 shows an example of a log file encoded using XML markup language in accordance with the invention.

Detailed Description of the Drawings

Fig. 1a shows the basic steps of a routine processing a continuous audio stream in accordance with the invention. After the routine is started 10 and the audio stream is digitized (not shown here) the digitized audio stream is analyzed in order to locate speaker changes 20. A lot of speaker change and speaker detection algorithms are known in the literature. For a comparison of techniques see for example F. Bimbot et al., Second-Order Statistical Measures for Text-Independent Speaker Identification, Speech Communication, Vol. 17, p.177-192, 1995. Hereby, for instance, the audio stream can be analyzed at frequency bands in order to derive characteristic features for different speakers. For a speaker change detection such feature vectors may be subjected to classical change detection techniques as described in the textbook by M. Basseville and Igor V. Nikiforov, Detection of Abrupt Changes: Theory and Applications, Prentice Hall Englewood Cliffs 1993, whereas for speaker identification the features are matched against a database of known speakers (S. Furui, An Overview of Speaker Recognition Technology, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, p.1-9, Martigny 1994).

If a speaker change is detected, for at least part of the following audio stream a speaker recognition is performed 30. Otherwise the speaker change detection is repeated until a speaker change is detected at all. After the speaker recognition 30 is finished it is checked 40 whether the recognized speaker is equal to a predetermined or preselected speaker or alternatively whether the speaker is recognized as a known speaker at all. If so at least the above mentioned part of the audio stream is transcribed, e.g. into plain text by means of a known speech recognition technique.

Now referring to the flow diagram depicted in Fig. 1b, a continuous audio signal 100, either recorded by means of an analog storage media or provided real-time, is first digitized

105. The digitized audio data are then scanned 110 whereby it is checked during loop 115 whether a speaker change occurs 120 and whether the detected new speaker is identical with a predefined or known speaker. The latter step is performed by means of speaker recognition 130 using prior art technology.

It is emphasized that the steps of detecting a speaker change 120 and performing a speaker recognition 130 can be alternatively performed in the reverse directon wherein the results of the speaker recognition 130 themselves can be used in order to detect speaker changes 120 thus simplifying the above described approach.

If the speaker change detection 120 reveals that a speaker change has occurred, the current time is taken 125 and protocolled e.g. in a log file. Having performed the speaker recognition 130 it is checked 135 whether the recognized speaker is identical with a predefined speaker. If true, the audio segments starting with the detected speaker changes are indexed 140 by using the protocolled time 125.

The scanning of the audio stream is continued 150 until the entire audio stream is scanned through and analyzed in the above described manner. Having finished the scan, only for the segments corresponding to selected speakers a speech recognition procedure, as known in the prior art, is performed 160. Hereby, in a preceding step 155, a speaker-related voice tract model and/or dictionary for the recognized speaker (step 130) is selected wherein the speech recognition 160 is performed based on that dictionary.

It is further noted that the steps 110 and 150 are optional and related to a scenario where an audio stream stored on a data carrier is scanned offline in order to perform the method according to the invention. Without these steps the mechanism can be performed for a real-time audio stream like a speech

signal incoming in a speech or voice recognition system.

Using a time base generator (step 102) as an external time reference for enabling writing of the time tags (step 140) is also optional and needed only in cases where the original audio signal does not comprise timing information.

The described method advantageously enables to perform speech recognition only for audio segments in a continuous audio stream which have been uttered by a given speaker.

Fig. 2 depicts a first embodiment of an apparatus according to the invention. In this embodiment, the continuous audio stream is recorded on a tape storage 200. First the audio stream is digitized by means of a prior art digitizer 210 particularly revealing digitized timer information 220 for the audio stream. In addition, the digitized audio stream is searched for speech/non-speech boundaries by means of an appropriate detector 230 also well-known in the prior art. The non-speech detector 230 delivers first candidates of speaker-change boundaries in form of first audio segments.

For these audio segments an utterance analysis is performed by means of an utterance analyzer & change detector 240. The audio stream is analyzed by an utterance analyzer which scans through the audio stream in order to gather speaker-specific audio features. For instance, the utterance analyzer can be implemented as a spectrum analyzer which takes information in the neighborhood of frequency bands which are characteristic for different speakers. The analyzed utterance signal is forwarded to an utterance changed detector which detects speaker changes. If an utterance or speaker change is detected by detector 240, the time of the speaker change is taken from the timer information provided by the digitizer 210, or an external timer, and written to a log file 255 stored in a database 260 by means of an indexer 250. It should be noted, that in many SRS systems,

the utterance analyzer is already an integrated part of the SRS (e.g. P.S. Gopalakrishnan et al, Acoustic models used in the IBM System for the ARPA HUB4 task, Proc. of the Speech Recognition Workshop, ARPA, 1996).

For most applications, the time information alternatively can be taken from the clock of a computer system or a dedicated hardware that is used to perform the speaker recognition. In cases where a higher precision is needed for the timing information, e.g. in an automatic logging or indexing of air control dialogues, the time can be taken from an external time reference that is merged with the audio stream during the digitization step.

Taking the logged index information together with the digitized audio stream provided by the digitizer 210, a speech recognition system (SRS) 270 as known in the prior art can perform a speech recognition procedure on the audio stream, but solely for the indexed audio segments.

It should be noted that the system described herein before processes audio data digitized by prior art technology. In e.g. a call center environment, such data are usually collected from the telephone set or the head set of an operator. For logging and archiving the digitized data stream is stored in a file, either on a call-by-call or shift-by-shift basis. The same digitized audio stream is then passed through the described speaker recognition system that computes features allowing the identification of individual speakers.

Now referring to Fig. 3, a second embodiment of the apparatus according to the invention is described. A real-time audio stream is input to a microphone 300 and digitized by means of a digitizer 310. The digitized audio stream is input to an utterance analyzer & change detector 320 in order to detect speaker changes as described above. A speech recognition system

(SRS) 330 has implemented a speaker model and/or dictionary change utility 340 which has access to different speaker-trained data 360, 370 stored in a database 350. Dependent on allegedly detected speaker changes, the dictionary change utility 340 can interchange between the different models 350, 370 thus providing an optimized multi-user SRS.

An example of a log file encoded using XML markup language in accordance with the invention is depicted in Fig. 4. The shown call center scenario starts with an incoming customer call 400 and a welcome text 410 spoken by an operator of the call center. The operator is assumed to be a preselected speaker with a known speaker ID which is „s0127" in the present example. Thus the start time and the end time of the welcome text 410 are marked with corresponding tags 420, respectively. The customer not being a preselected speaker with an ID is responding to the welcome text 410 and the audio signal tagged with the corresponding start time and end time accordingly but storing that the speaker ID is „unknown". Next the operator asks the customer for the customer number 440 wherein the audio signal is tagged again 450 with the known speaker ID. These steps are continued accordingly until the end of the call wherein in step 460 the operator repeats the customer number named by the customer in the preceding step and confirms the correct database entry of the customer „and the address is Helga Mustermann the ...". At the end of the call the audio signal is tagged with the endcall time 470.

It should further be noted that the above described method and apparatus can either be implemented in hardware, software or a combination thereof.

EPO - Munich
3
30. Nov. 2000

C L A I M S

1. A method of processing a continuous audio stream containing human speech related to at least one particular transaction, comprising the steps of:

digitizing the continuous audio stream;

detecting a speaker change in the digitized audio stream;

performing a speaker recognition if a speaker change is detected;

transcribing at least part of the continuous audio stream if a predetermined speaker is recognized.

2. A method of processing a continuous audio stream containing human speech related to at least one particular transaction, comprising the steps of:

digitizing the continuous audio stream;

detecting a speaker change in the digitized audio stream;

performing a speaker recognition if a speaker change is detected;

indexing the audio stream with respect to the detected speaker change if a predetermined speaker is recognized.

3. Method according to claim 1 or 2, comprising the further step of protocolling time information for detected speaker changes.

4. Method according to any of the preceding claims, wherein the step of detecting a speaker change and/or the step of performing a speaker recognition is/are preceded by the further step of detecting non-speech boundaries between continuous speech segments.

5. Method according to any of the preceding claims, wherein the step of detecting a speaker change is accomplished by use of at least one characteristic audio feature, in particular features derived from the spectrum of the audio signal.

6. Method according to claim 1 or 2, wherein the step of performing a speaker recognition involves the particular steps of calculating a speaker signature from the audio stream and comparing the calculated speaker signature with at least one known speaker signature.

7. Method according to any of the preceding claims for use in a speech recognition or voice control system comprising at least two speaker-specific speaker models and/or dictionaries, wherein interchanging between the at least two speaker-specific dictionaries dependent on the detected speaker change and the corresponding recognized speaker.

8. Apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction, comprising:

means for predetermining at least one speaker;

means for detecting speaker changes in the audio stream;

means for recognizing the predetermined speaker in the audio stream;

means for initiating transcription of at least part of the audio stream in case of a detected speaker change and a recognized predetermined speaker.

9. Apparatus for processing a continuous audio stream containing human speech related to at least one particular transaction, comprising:

means for predetermining at least one speaker;

means for detecting speaker changes in the audio stream;

means for recognizing the predetermined speaker in the audio stream;

means for indexing the audio stream dependent on a detected speaker change and a recognized predetermined speaker.

10. Apparatus according to claim 8 or 9, further comprising

means for detecting non-speech boundaries between continuous speech segments.

11. Apparatus according to any of claims 8 to 10, further comprising

means for automatically scanning a continuous audio record, in particular a continuous audio stream recorded on a data or a signal carrier, and for

detecting speaker changes in the continuous audio record.

12. Apparatus according to any of claims 8 to 11, further comprising

means for continuously monitoring a real-time continuous audio stream and performing the steps of claim 1 or 2.

13. Apparatus according to any of claims 8 to 12, further comprising log means for protocolling time information for the at least one detected speaker change.

14. Apparatus according to any of claims 8 to 13, comprising means for marking at least the beginning of a detected speech segment related to a predetermined speaker.

15. Apparatus according to any of claims 8 to 14, comprising data base means for storing speech signatures for at least two speakers.

16. Speech recognition or voice control system processing an incoming audio stream and having at least two speaker models and/or speaker-specific dictionaries, comprising

means for detecting a speaker change in the incoming audio stream;

means for gathering speaker-specific information and for comparing the gathered speaker-specific information with corresponding speaker-specific information of at least one predetermined speaker thus recognizing the at least one predetermined speaker;

means for interchanging between the at least two speaker-specific dictionaries dependent on the detected speaker change and the corresponding recognized speaker.

17. A data processing program for execution in a data processing system comprising software code portions for performing a method according to any of claims 1 to 7 when said program is run on said computer.

18. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform a method according to any claims 1 to 7 when said program is run on said computer.

EPO - Munich
3
30. Nov. 2000

A B S T R A C T

Disclosed are a method and apparatus for processing a continuous audio stream containing human speech in order to locate a particular speech-based transaction in the audio stream, applying both known speaker recognition and speech recognition techniques. Hereby it is enabled that only the utterances of a particular predetermined speaker are transcribed thus providing an index and a summary of the underlying dialogue(s).

In a first scenario, an incoming audio stream, e.g. a speech call from outside, is scanned in order to detect audio segments of the predetermined speaker. These audio segments are then indexed and only the indexed segments are transcribed into spoken or written language. Thus an already occurred specific transaction can be found on an endless storage media like a magnetic tape. The proposed mechanism thus makes the task of locating an audio log of a specific transaction a much more less effort.

In a second scenario, two or more speakers located in one room are using a multi-user speech recognition system (SRS). For each user there exists, a different speaker model and optionally a different dictionary or vocabulary of words already known or trained by the speech or voice recognition system. In such an environment, the invention allows to switch between different dictionaries when a first user has stopped utterance and a second user is going to start his utterance.
(Fig. 1B)

EPO - Munich
3
30. Nov. 2000

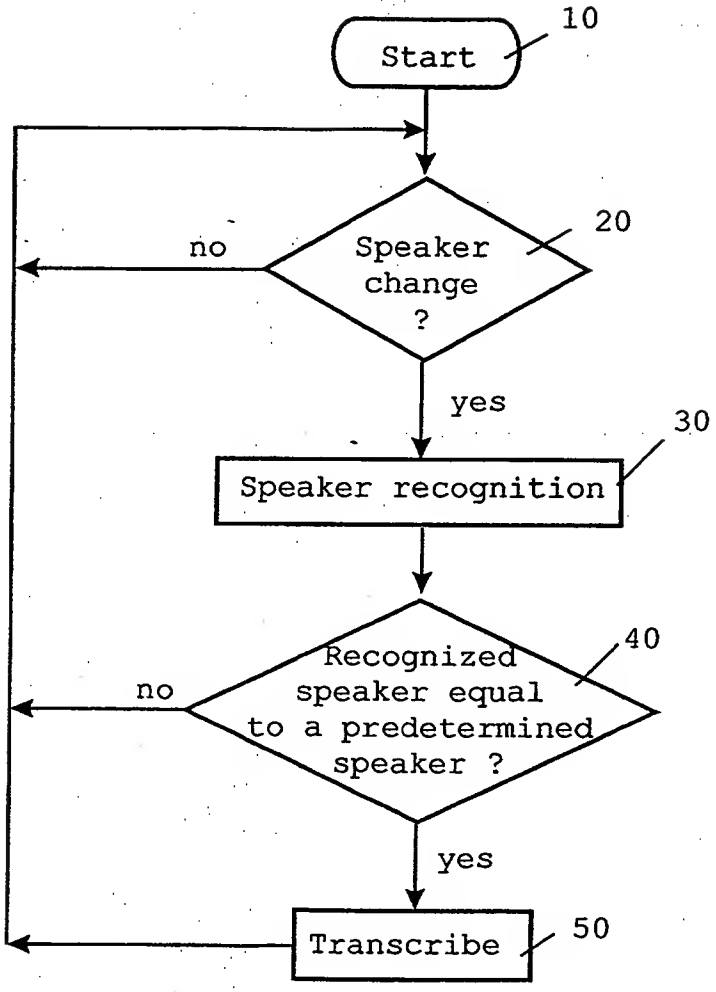

FIG. 1A

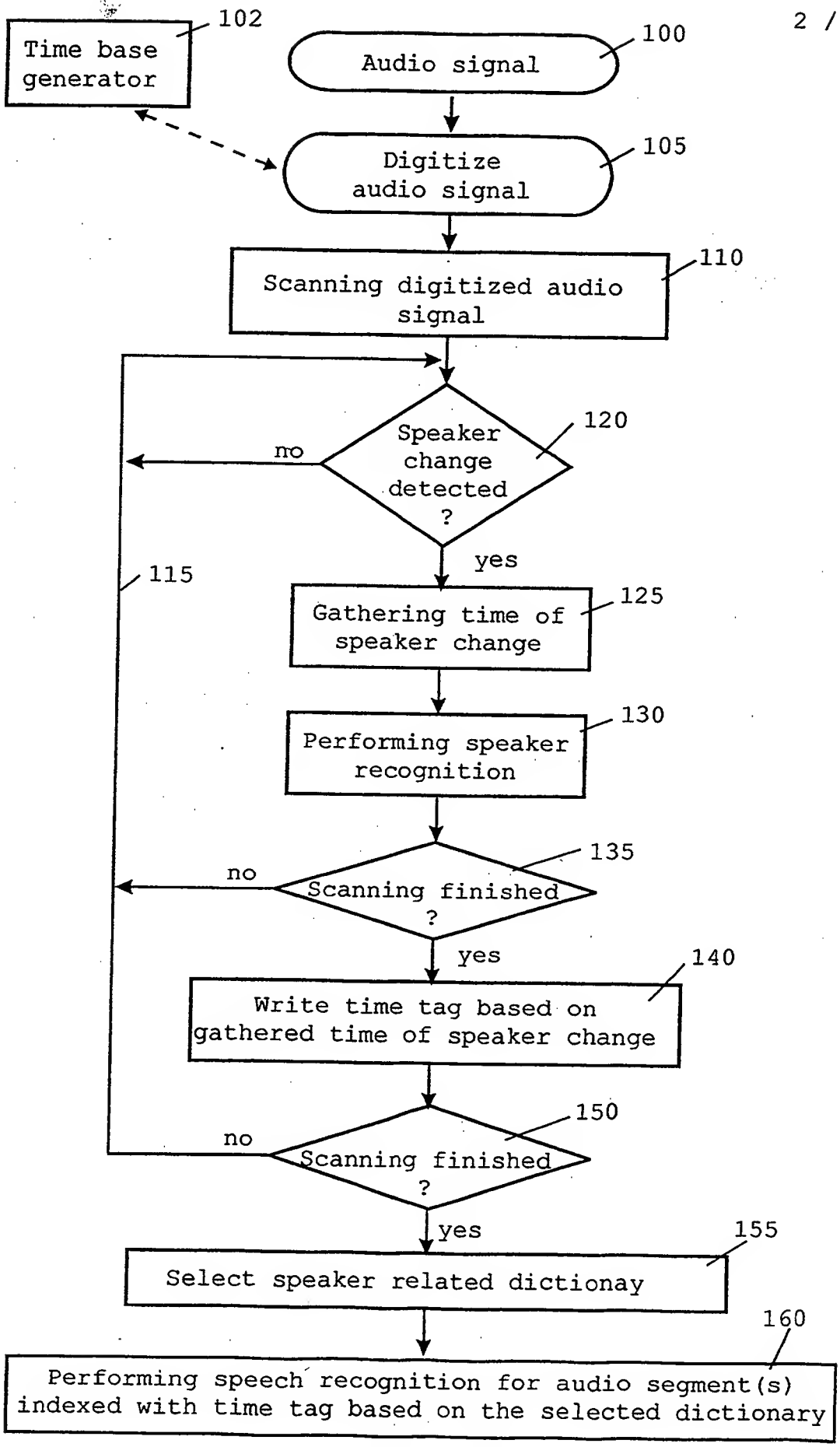
Time base generator
102
Audio signal
100
Digitize audio signal
105
Scanning digitized audio signal
110
Speaker change detected ?
120
no
yes
115
Gathering time of speaker change
125
Performing speaker recognition
130
Scanning finished ?
135
no
yes
Write time tag based on gathered time of speaker change
140
Scanning finished ?
150
no
yes
Select speaker related dictionay
155
Performing speech recognition for audio segment(s) indexed with time tag based on the selected dictionary
160

FIG. 1B

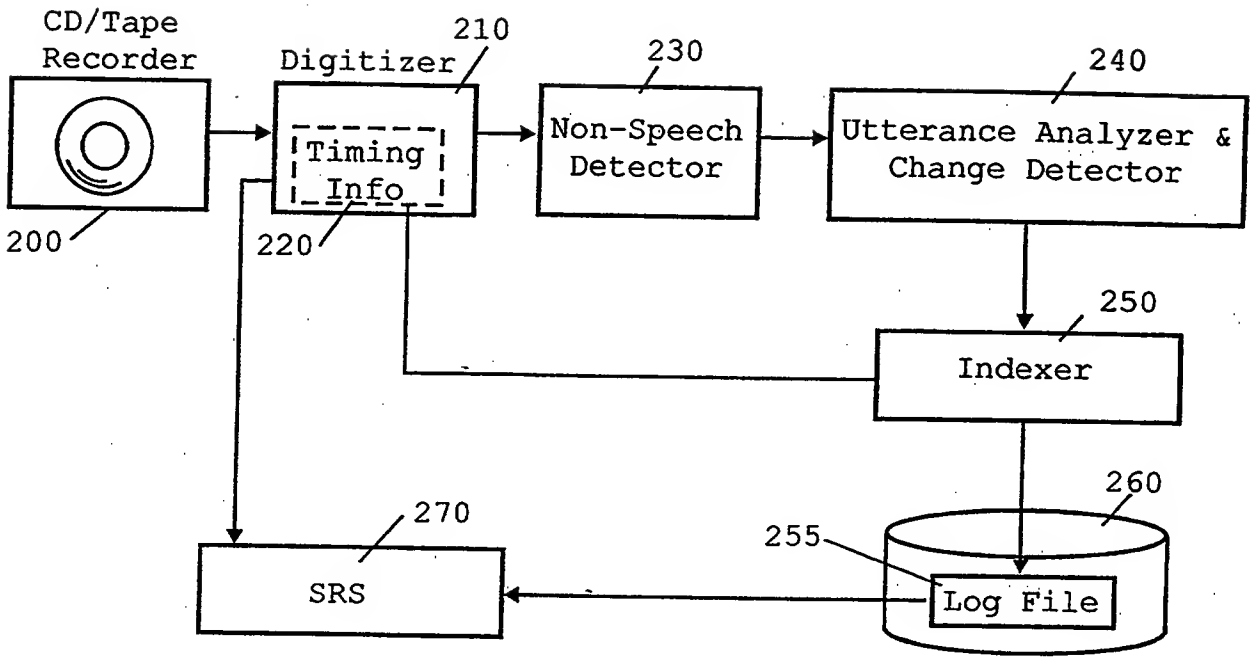
CD/Tape Recorder
200
Digitizer
210
Timing Info
220
Non-Speech Detector
230
Utterance Analyzer & Change Detector
240
Indexer
250
Log File
255
260
SRS
270

FIG. 2

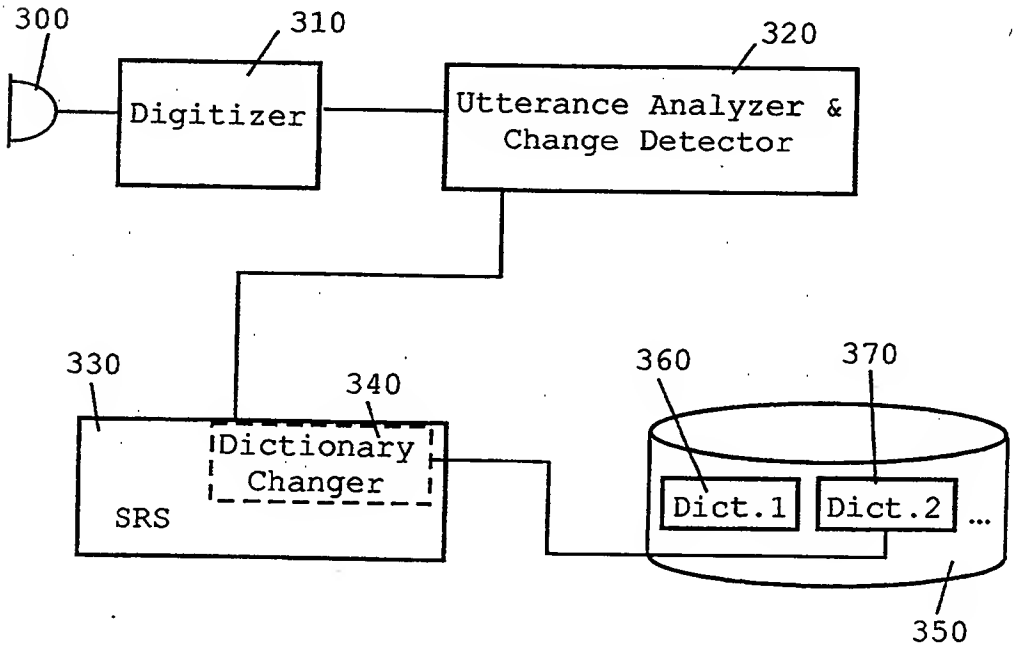
300
Digitizer
310
Utterance Analyzer & Change Detector
320
330
SRS
340
Dictionary Changer
360
Dict.1
370
Dict.2
...
350

FIG. 3

```
<transcript file="audio/20000510.log">
...
<begincall time="08:00:04:30"/>
<speaker id="s0127" starttime="08:00:04.30"
endtime="08:00:14:40">
<text>Satisfaction Guaranteed Phone Order
Department</text>
</speaker>
<speaker id="unknown" starttime="08:00:17.40"
endtime="08:00:30.30"/>
<speaker id="s0127" starttime="08:00:30.30"
endtime="08:00:34:00">
<text>To process your order I need your customer
number please</text>
<speaker id="unknown" starttime="08:00:40.30"
endtime="08:03:30.30"/>
<speaker id="s0127" starttime="08:03:30.30"
endtime="08:03:34:00">
<text>137532 and the address is Helga Mustermann,
Birkenweg 7, 12345 Dorf right</text>
<speaker id="unknown" starttime="08:03:04.30"
enditme="08:06:06.00"/>
...
<endcall time="08:10:25.01"/>
</transcript>
```

400 420 410 430 450 440 460 470

FIG. 4